Analyze the Future

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

WHITE PAPER

Itanium 2–Based Servers: Deployments Shipping Across Form Factors and Price Points

Sponsored by: ISA

Matthew Eastwood Jean S. Bozman

October 2006

EXECUTIVE SUMMARY

Itanium 2–based servers are shipping in a variety of form factors, across a range of price points, and are being deployed for a spectrum of workloads. This variety is not widely understood in the industry due to an initial focus on the most scalable Itanium-based servers. As the ecosystem has matured, more than 8,500 applications have become available for Itanium-based servers — and deployments range from scale-out volume servers in server farms and clusters to midrange enterprise servers and high-end enterprise servers in large corporate datacenters.

IDC supply-side data shows that EPIC (Itanium-based) servers generated $2.3 billion in customer revenue in 2005, up from $1.4 billion in 2004. Customer revenue is defined as the combination of factory revenue and channel uplift through the addition of peripheral devices shipped to customers. More than half of the 2005 shipments were rack-optimized servers in the volume or midrange enterprise server class. Looking forward, IDC estimates that $6.8 billion in customer revenue for Itanium-based servers will be shipped in 2010, showing a five-year CAGR of 23.6%.

Workloads for these servers were varied as well, including business processing (30% of EPIC server revenue), decision support (21.2%), application development (9.6%), and collaborative (6.5%) workloads — in addition to IT infrastructure, Web infrastructure, and technical workloads. The workloads data is based on an IDC 2005 customer survey of more than 900 IT managers. Respondents, who had responsibility for managing IT operations, described all workload types. However, the high-performance computing (HPC) or technical workloads may be somewhat underrepresented in this sample of central-site computing professionals due to the deeper focus on HPC and technical workloads within business units, departments, and divisions of enterprises and organizations.

Itanium-based servers are being deployed in more and different ways from when they were first made available in 2002. As the ecosystem for Itanium-based servers matures, the deployments are showing more variety, both in terms of the physical servers deployed and in terms of the workloads they run. As a result, Itanium-based servers can be mapped across the computing spectrum and can be configured for both scale-out and scale-up server requirements.

INTRODUCTION

A Range of Servers, a Spectrum of Use

The worldwide server market supports a remarkable range of servers, including many price points, many operating systems, and many form factors. Indeed, computing capability can be delivered in a number of ways: in partitions of large servers, in separate rack-optimized servers, and on blades within blade-server chassis.

The reason for this variety is straightforward: There is a broad computing spectrum — and an equally broad spectrum of customer preferences for the style in which processing power is packaged, deployed, and maintained.

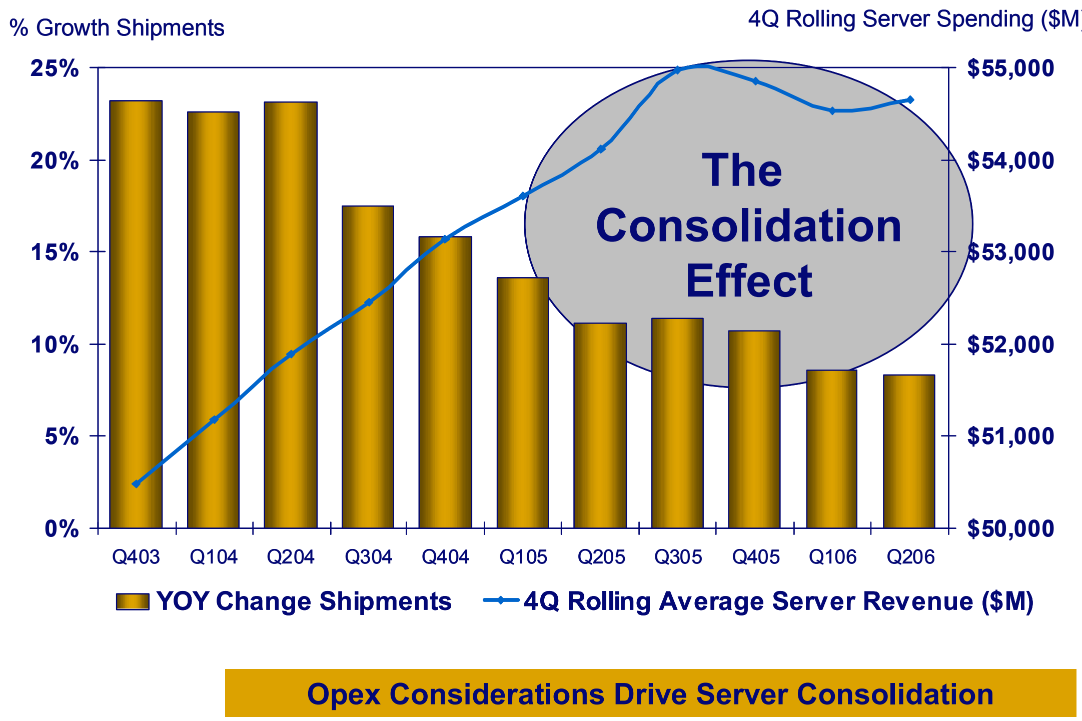
In recent years, there has been rapid growth in low-cost servers as IT customers have sought to contain capital expenditures (capex), even as there was continued strength in scalable server deployments that consolidated workloads and in projects focused on controlling operational expenditures (opex).

Today, the focus centers on opex, given the issues associated with power/cooling, server utilization, and efficient resource management that arise when workloads are scattered among very large scale-out deployments without adequate provisioning, workload management, and system management software being put into place.

IDC server data shows that IT customers are shopping for value, which has led to richer configurations of volume servers, a slower rate of growth in server unit shipments worldwide, and continued use of midrange enterprise and high-end enterprise servers within the corporate network to run the most mission-critical workloads (see Figure 1). Both midrange and high-end systems are used as consolidation platforms for the most demanding types of business processing workloads. This analysis clearly shows that IT spending patterns are shifting, as opex-driven alternative platforms for IT consolidation represent a significant consideration for new server purchases around the world.

FIGURE 1

Opex Considerations Drive Scalability and Consolidation



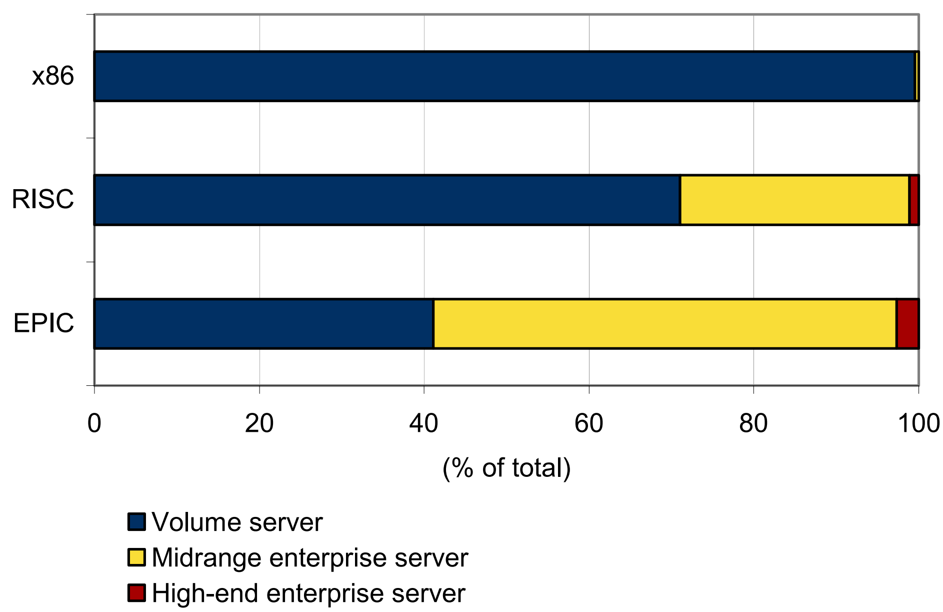
Source: IDC, 2006

While opex considerations are clearly becoming more top of mind during the server procurement process, the number of server platform options available to IT customers remains significant. For example, within each server architecture platform, as defined by the processor type on which it is based, is a range of form factors and price points. This is true for x86 servers, RISC servers, and EPIC (Itanium 2–based) servers. IDC worldwide server data demonstrates this range of server shipments within each of these categories, although the mix of servers when viewed by form factor and price category is different.

Figure 2 provides a view of server shipments by processor architecture and server class with volume systems priced below $25,000; midrange enterprise servers ranging from $25,000 to $499,000; and high-end enterprise servers priced at $500,000 or more. More than half of all spending for EPIC or Itanium-based systems is for midrange enterprise systems — and the rest of the spending spans a range of price points, depending on the number of sockets per server, and on the richness of the configuration, as shipped by the server vendor. These systems typically have between 4 and 16 sockets, and the typical processor is undergoing a transformation from single-core to dual-core technology, which further increases the processing throughput that each of these server systems is capable of delivering. The increased core density, along with the advanced software functionality on midrange enterprise systems, has allowed midrange enterprise servers to take on some of the roles that were historically associated with high-end servers alone. Today, there is more of a blend between high-end midrange enterprise servers and the rest of the high-end enterprise server segment, with the category for each named server reflecting the price point of the server being shipped — above or below $500,000.

FIGURE 2

Worldwide Server Shipments by Server Class and CPU Type, 2005



Source: IDC, 2006

The idea that Itanium-based servers can take on workloads formerly hosted by mainframes or by RISC-based servers is clear. Many server vendors that offer Itanium-based servers today have offered traditional mainframe and RISC server platforms in previous generations of server technology. Therefore, this knowledge base has been leveraged by these vendors and "baked into" the latest generation of scalable Itanium-based servers in the midrange enterprise and high-end enterprise server segments.

Indeed, these high-end EPIC servers carry many features that originally came out of the mainframe datacenter, including partitioning, virtualization, resource reallocation, workload balancing, and clustering of datacenter servers, to reach extremely high levels of availability for applications and data. In addition, compared with the price points for older generations of high-end datacenter server systems, the price points for these high-end servers are competitive due to their new technology base of off-the-shelf components and the leverage of industry-standard memory and I/O devices.

However, as Figure 2 clearly shows, a wide range of rack-optimized servers in both the volume and midrange enterprise segments is selling into the marketplace from a variety of vendors worldwide. This customer-based data shows that EPIC/Itanium servers have a range of shipment types across the entire computing spectrum — volume, midrange enterprise, and high-end enterprise — not just high end. The variety of form factors, including many rack-optimized servers, demonstrates the broadening of the Itanium ecosystem as more Itanium 2–based servers ship into the worldwide marketplace.

If rack-optimized servers are becoming such a widely deployed form factor, what does that mean for Itanium-based server deployments? What are customers doing with these configurations? And, equally important, what workloads and vertical markets are these rack-optimized servers supporting? The next section of this paper examines these questions by providing a close-up look at the Itanium-based server data by server-class segment, by workload, and by operating system type. Vertical market examples are also provided.

IDC WORLDWIDE SERVER MARKET DATA

Itanium-based servers are being deployed across a wider range of usage scenarios than has been perceived in the market. In general, Itanium-based systems have been positioned as alternatives or replacements for mainframe or RISC-based servers. However, when the worldwide server market data is considered, inclusive of its price bands and segmentation, more features of the EPIC/Itanium 2–based server market can be seen.

IDC server data clearly shows that most enterprises continue to deploy a mix of both small and large servers in a variety of form factors — pedestal (non-rack-optimized) systems, rack-optimized servers, and bladed architectures. But in the case of EPIC systems, these server form factors are being deployed to perform a broader range of computing tasks or workloads than has been appreciated. These enterprise and HPC workloads remain critical to business operations, but the servers themselves are often deployed in a variety of ways.

Looking more closely at the EPIC architecture platform, which is based on Itanium 2 processors, we can clearly see the range of computing power shipped to IT customers. Although many industry observers may perceive that only large, scalable Itanium-based servers are shipping, IDC data shows that volume servers and midrange enterprise servers are central to overall shipment patterns. In fact, in 1Q06, rack-optimized volume servers accounted for 40% of all units shipped and rack-optimized midrange servers accounted for 25% of all units shipped.

USAGE SCENARIOS

Itanium-based servers are often seen as high-end servers, as mainframe replacements, and as RISC server replacements. They fit all three of these roles when deployed in large datacenters. But the deployments of volume and midrange servers include a number of different usage patterns for Itanium-based servers, including the following:

- **Application development platform.** Itanium-based servers are used as platforms for application development — whether for Web-enabled applications, creation of Web portals, custom applications for specific vertical markets, or development of ISV software packages.

- **Test/development platform.** Support for partitioning within Itanium-based servers allows for physical isolation of workloads so that test/development workloads can be run side by side with production workloads without interference. This capability is important for continuous support of mission-critical workloads and for testing new applications for eventual deployment.

- **Node for HPC workloads.** Large research organizations, technical centers within manufacturing companies, and datacenters supporting financial services companies have all deployed HPC clusters built with Itanium-based server nodes.

- **Node for enterprise clustered-server configurations.** Scale-out deployments of Itanium-based servers include those designed to support failover high-availability clusters, load-balancing clusters, scalability clusters, and management clusters. Some have been deployed as resources within HPC or enterprise grids.

- **Database platform and data warehouse platform.** Database and database-related workloads are key components of the business processing that takes place on Itanium 2–based servers. These systems leverage 64-bit processing to speed database performance, moving large data sets into and out of memory — and directly addressing multigigabyte segments of the database at once.

- **Business intelligence (BI) platform.** Business intelligence and analysis of large data sets to discover trends within the data benefit from 64-bit processing, support for parallelized instruction processing, and support for consolidation of growing databases.

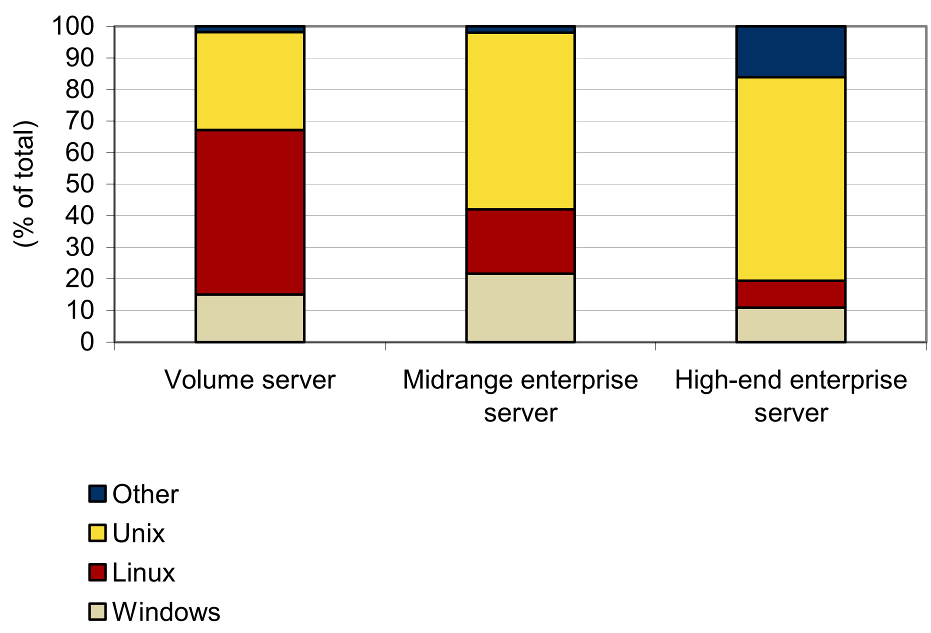
Further variation is seen by workload type: Although many initial shipments of Itanium-based servers supported Unix and Linux workloads, including those in HPC and technical computing clusters, in recent years, Itanium-based servers running Microsoft Windows have gained more technical and enterprise workloads than was the case in 2002–2004.

The EPIC/Itanium architecture is supported by a number of leading systems suppliers, including Fujitsu Limited, Fujitsu Siemens Computers, Groupe Bull, Hitachi Ltd., HP, NEC, SGI, and Unisys, which together accounted for 95% of all Itanium-based server revenue in 2005. While these companies are the segment leaders by factory revenue, other vendors also participate in the wider Itanium-based server ecosystem, including smaller, regional systems vendors as well as software vendors, storage vendors, services vendors, value-added resellers (VARs), and systems integrators (SIs). Thus, Itanium-based server sales are made both directly (by vendors' sales forces) and indirectly (via vendors' channel partners). In 2005, $2.3 billion was spent on Itanium-based servers worldwide in terms of customer revenue.

Figure 3 shows that the Itanium-based systems market is driven by applications deployed in a number of different operating environments. In addition, IDC sees variation in Itanium-based server shipments by server workloads and by vertical market focus.

FIGURE 3

Worldwide Server Shipments by Server Class and Operating System, 2005



Source: IDC, 2006

From the beginning, enterprise customers have been attracted to the Itanium 2 architecture in part by the broad operating system support on the platform. In 2005, more than 35% of all Itanium 2 processor spending was for Windows and Linux systems in support of a wide range of enterprise and HPC workloads. IDC believes that these mission-critical enterprise-class workloads include a variety of business processing applications such as finance, accounting, human resources, supply chain, and other ERP workloads.

In addition, the reliability, availability, and serviceability (RAS) features available on the Itanium 2 processor are attractive to customers looking for a platform for enterprise database workloads including data warehousing and real-time business analytics, such as data analysis/data mining and CRM. These RAS features of Itanium 2 processors include:

- Enhanced Machine Check Architecture (EMCA), with well-defined interfaces for error handling at the hardware, firmware, and operating system levels

- Error correction code (ECC), which corrects any transient errors in the code itself that could cause processing results to become incorrect due to corrupted data

- Compilers from Intel, Microsoft, and others that help developers to optimize code for running on the parallelized Itanium 2 architecture and on the Dual-Core Itanium 2 processor 9000 series now shipping in vendor server systems

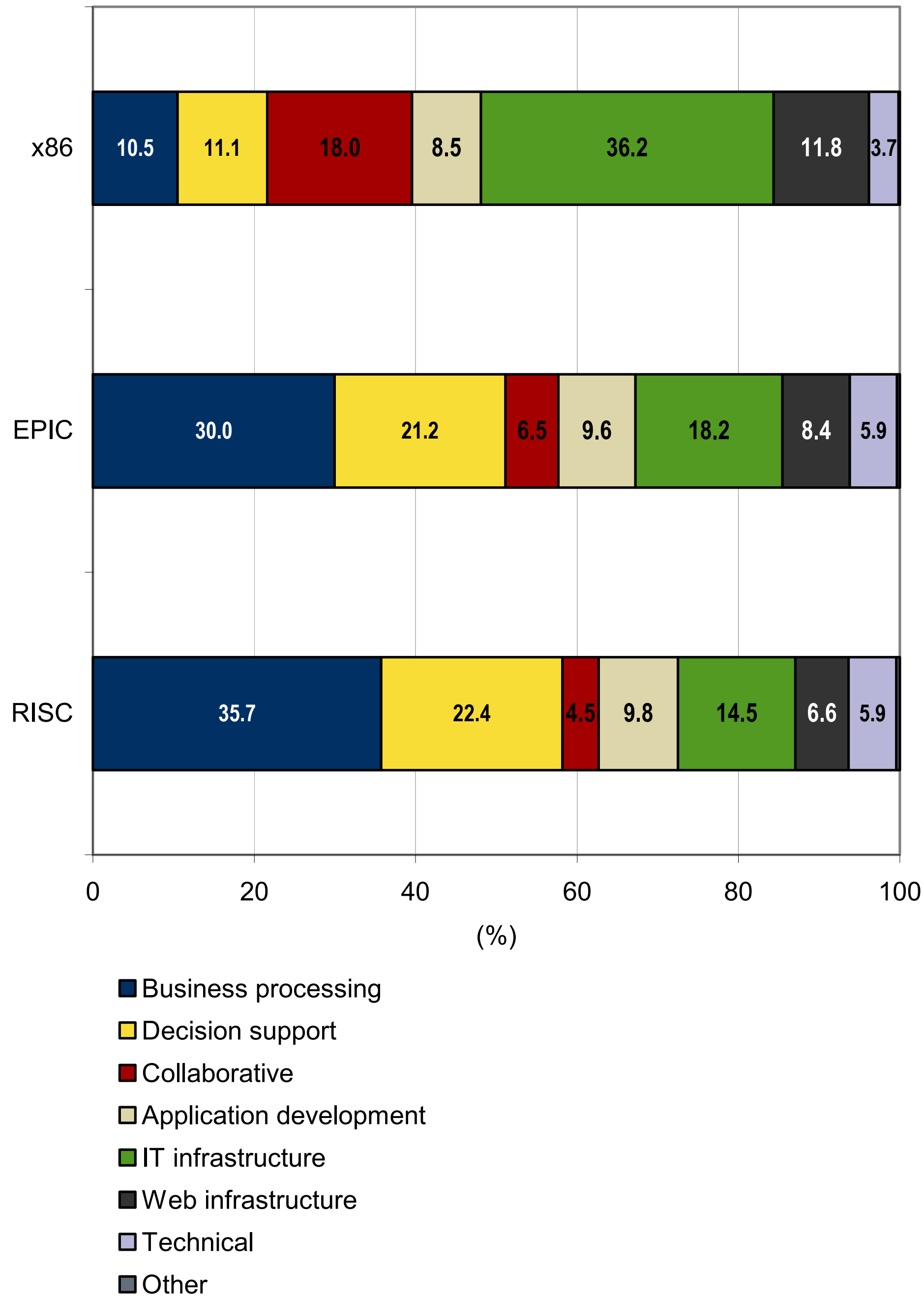
- Intel development tools for programming to Itanium 2 processors, including the Intel Thread Checker, Intel Thread Profiler, and Intel Threading Building Blocks, all of which are designed to make it easier for software developers to realize significant performance gains with multicore and Itanium 2 architecture

SERVER WORKLOADS DATA

IDC has long believed that a strategic fit exists between server platform and workload, a linkage that has been demonstrated over six years of IDC workloads research. In fact, as Figure 4 shows, the server workload or application type is clearly linked to the underlying processor architecture of the server.

FIGURE 4

Worldwide Server Customer Revenue Share by Workload and CPU Type, 2005



Source: IDC, 2006

For Itanium-based systems, more than half of all server spending is driven by business processing and decision support workloads. These workloads typically require platforms for which a heavy engineering emphasis was placed on RAS features. This mix of workloads reflects a strong emphasis on business processing and mission-critical workloads.

In addition, the characteristic of scalability is often a requirement for the target workload because scale-out deployments are not always sufficient to deliver the level of throughput performance and single-system image (SSI) support that is necessary for these mission-critical workloads.

IDC defines a series of workload types, including business processing (e.g., OLTP, BI, and ERP), decision support (typically database-related analytical workloads), collaborative (e.g., email or groupware), application development, IT infrastructure (e.g., network protocol, file/print), Web infrastructure (e.g., Web serving, proxy, and caching), technical (e.g., HPC, which includes scientific and engineering workloads), and other (refer back to Figure 4). The IDC workloads studies on which this type of analysis is based have been conducted each year since 1999, and the results have been widely published.

COMPUTING ACROSS MULTIPLE TIERS

A wide array of deployment types was seen within the workloads data, with thin, dense rack-optimized servers installed into the Web tier and application tier of computing deployments, while more scalable servers were seen in the application-serving and database-serving tiers of IT infrastructure. Usage scenarios that shed light on solution types associated with server workloads include:

- **Database engines.** Support for a variety of databases, including Oracle10g and Oracle 9i RAC clustered configurations; Microsoft SQL Server 2005; IBM DB2 for Windows and Linux; MySQL, Progress, and Postgres, and other relational databases. The Itanium 2 architecture 64-bit design, and its requirement that 64-bit operating systems be used, means that it can leverage 64-bit addressing to directly address large data sets associated with fast database scans, decision support analysis, and BI analysis. That analysis is based on data stored within the underlying database product that was built up, over time, on the basis of transactional data associated with ongoing business or ongoing operations within an organization.

- **ERP platforms.** Customers have already deployed ERP software (e.g., SAP and Oracle) on Itanium-based servers. Consolidation of workloads that formerly ran on many modules, scattered across the enterprise, is also prompting some amount of consolidation of ERP workloads on Itanium-based servers, as this trend also affects other types of server platforms.

- **Business intelligence.** The large size of the databases that underlie BI workloads requires scalable resources and high levels of availability for applications and data. Itanium-based servers are one of the platforms considered by IT customers who are planning new BI consolidation projects.

- **HPC clustered compute nodes.** Since the first wave of Itanium processor shipments in 2001 and 2002, Itanium-based servers have been deployed as compute nodes, especially in clusters of servers running the Unix and Linux operating environments. Often, these HPC clusters leverage open source software, such as MPI and Beowulf, to orchestrate the workflow among the server nodes. Examples can be seen in seismic analysis for oil/gas industry applications and computational fluid dynamics (CFD), as leveraged by aerospace and automotive companies analyzing airflow over their prototypes and products.

- **Telecommunications servers.** The telecommunications world leverages Unix and Linux operating systems to run network switches, billing systems, telco security, telecommunications middleware, and even cellular phone call-pattern analytics. Many next-wave deployments that support IP-enabled traffic (e.g., IPTV, VoIP) are also expected to run on Itanium-based servers.

- **Support for search engine functionality.** An emerging area for Itanium-based servers is as an engine for search engine platforms. The Itanium 2 processor's built-in support for parallelism in workloads and RAS features are factors in the consideration of Itanium-based servers in this new area of IT infrastructure deployment.

- **Support for vertical-market applications.** As the inventory of off-the-shelf ISV packaged applications tops 8,500 titles (as of mid-2006), Itanium-based servers can take on a wider range of vertical workloads, depending on the availability of a given ISV package for the Itanium-based platform. Examples of verticals include financial services, retail, manufacturing, telecommunications, and logistics/ distribution.

CHALLENGES AND OPPORTUNITIES

Itanium-based server providers face the perception that their servers are not widely used and that they do not run a variety of workloads. Indeed, the perception that they are only mainframe replacements does not acknowledge the actual shipments of many rack-optimized servers into a broader range of server deployments, including both the volume and midrange enterprise server market segments.

This broader deployment is one reason that Itanium-based server factory revenue reached $2.2 billion in 2005, which is 1.6 times the $1.38 billion in Itanium-based server factory revenue in 2004 and 4.6 times the $480 million in Itanium-based server factory revenue in 2003.

Itanium-based servers will continue to compete with servers based on other hardware architectures as customers continue to evaluate new server platforms for next-generation IT infrastructure. The competition includes x86 servers based on Intel or AMD processors, RISC servers, and traditional CMOS host servers for the datacenter, including mainframes. However, within that larger competition arena are multiple market segments — form factor, operating system, configuration, and vertical market — where Itanium-based servers can be considered as potential replacements to existing hardware platforms.

The opportunity is for vendors of Itanium-based servers, working in concert with ISVs, VARs, and SIs, to discuss deployment patterns among the customer base — and to do so in an educational, descriptive way. If a wider segment of the IT customer community understands that Itanium-based servers support a wider range of price points, deployment styles, workloads, and usage patterns, then it could prompt further evaluations of Itanium-based platforms for new projects.

CONCLUSION

Itanium-based servers are shipping into a variety of market segments today. Although it is widely known that Itanium-based servers can take on workloads that formerly were destined to run on mainframes or RISC-based servers, their role as rack-optimized servers in smaller form factors is not as well known.

IDC server data demonstrates the range of Itanium-based servers on a worldwide basis. More than half of Itanium-based servers shipped in 2005 were rack-optimized — and most of those servers were shipped in volume server form factors, with one or two sockets. Both volume server and midrange enterprise servers can be deployed as nodes for commercial or HPC clusters, while midrange and high-end enterprise servers can also act as platforms for workload consolidation.

In summary, the Itanium-based server space features a wide array of deployment types — spanning multiple tiers of computing, workload types, and form factors — as they are delivered into the worldwide server marketplace.